Document de travail

—

Méthode de vérification statistique pour l'Enquête sur le Commerce de Gros et le Commerce de Détail

Jean-Marie Berthelot
D.M.E.E.
Mars, 1985

TABLE DES MATIÈRES

Page

I. Introduction 1

II. Description théorique de la vérification statistique 1

III. Utilisation des paramètres A, c et u 8

IV. Propriétés de la méthode proposée pour la vérification statistique .. 12

V. Bibliographie 13

Appendice

I. Test sur des données réelles

1. Introduction

On demande aux compagnies échantillonnées dans le cadre de l'enquête sur le Commerce de Gros et le Commerce de Détail (ECGCD) de rapporter mensuellement leurs ventes et leurs stocks. Or on observe que les données rapportées par certaines compagnies présentent des fluctuations mensuelles ou annuelles importantes comparativement aux autres compagnies oeuvrant dans le même secteur économique. De telles compagnies sont appelées observations aberrantes et doivent faire l'objet d'une attention spéciale pour établir la validité de l'information rapportée. La vérification statistique est la technique qui permet d'identifier les observations aberrantes en utilisant de l'information statistique obtenue à partir des données de l'enquête.

Le but du présent document est de décrire la méthode suggérée pour la vérification statistique des données.

2. Description théorique de la vérification statistique

Soient $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, n paires d'observations pour une donnée financière (ventes ou stocks)

où x_i, i=1, 2, ..., n dénote une valeur courante,

et y_i, i=1, 2, ..., n dénote une valeur précédente

Soit r_i, le rapport entre la valeur courante et la valeur précédente pour l'observation i.

Exemple: Soit $x_1 = 1$, $y_1 = 3$, $x_2 = 3$ et $y_2 = 1$

Alors $r_1 = 1/3$ et $r_2 = 3$

On note que même si r_1 et r_2 ont des valeurs différentes, ils représentent la même magnitude de changement. Le rapport $r_1 = 1/3$ représente une diminution par un facteur de 3 tandis que le rapport $r_2 = 3$ représente une augmentation par un facteur de 3.

Afin d'avoir une vérification statistique qui traite de la même façon une augmentation et une diminution, la transformation suivante est appliquée aux rapports r_i.

$$S_i = \begin{cases} 1 - 1/r_i & \text{si } r_i < 1 \\ r_i - 1 & \text{si } r_i \geq 1 \end{cases}$$

Exemple:

1) $r_1 = 1/3$ implique $S_1 = -2$
2) $r_2 = 3$ implique $S_2 = +2$

La transformation "S" associe une valeur identique à des changements de même magnitude et utilise le signe (+/-) pour représenter la direction de ces changements. Les valeurs S_1 et S_2 générées par l'exemple ci-dessus, associent une valeur de "2" à chacun des changements, le signe (+/-) indique si le changement est une augmentation (+) ou une diminution (-). La transformation "S" introduit une symmétrie autour de zéro. L'introduction d'une symmétrie autour de zéro avec des valeurs positives associées aux augmentations ~~et des~~

~~valeurs positives associées aux augmentations~~ et des valeurs négatives associées aux diminutions, facilite la construction d'une procédure de vérification statistique qui traite de la même façon, une augmentation ou une diminution d'une donnée financière.

Pour l'ECGCD, non seulement le changement associé à une donnée financière est important, mais aussi la taille de la donnée financière associée à ce changement. Afin de tenir compte de la taille d'une unité, la transformation suivante est appliquée aux S_i:

$$e_i = S_i * (x_i ** u)$$

où $0 \leq u \leq 1$

Les e_i sont appelés "effets". Le terme effet est utilisé, puisqu'un e_i représente une mesure relative de l'effet de la variation de la $i^{ième}$ unité sur l'estimation totale de la donnée financière x.

La valeur de l'exposant "u" dans la transformation "E" permet de contrôler l'importance qui va être associée à la valeur de la donnée financière. Il est bon de noter qu'une valeur de u=0 implique l'égalité entre les S_i et les e_i.

Exemple:

Supposons que nous avons 3 unités dans l'enquête:

1) x_1=3,000, y_1=1,000 ce qui implique: r_1=3 et S_1=2

2) x_2=30,000, y_2=10,000 ce qui implique: r_2=3 et S_2=2

3) x_3=300,000, y_3=100,000 ce qui implique: r_3=3 et S_3=2

En utilisant une valeur de u=0.5, les valeurs des e_i sont calculés pour ces 3 unités de la façon suivante:

e_1 = 2* (3,000**0.5) = 109.5

e_2 = 2* (30,000**0.5) = 346.4

e_3 = 2* (300,000**0.5) = 1,095.4

Cet exemple montre que pour 3 unités qui ont la même tendance (r=3), une valeur de e différente est associée à chacune des unités et que la valeur de e augmente quand la valeur de la donnée financière augmente.

Pour l'ECGCD, la vérification statistique consiste à identifier les unités ayant une valeur de "e" aberrante, en utilisant la méthode de détection par les quartiles. La méthode de détection par les quartiles peut être décrite de la façon suivante:

Soit e_i, l'effet associé à l'unité i et faisons l'hypothèse que les valeurs des e_i sont par ordre croissant, c'est-à-dire:

$$e_1 \leq e_2 \leq \cdots \leq e_n$$

Définissons W_i, le poids associé à l'unité i, comme étant 1/n, $\forall$ i. Ce qui implique:

$$\sum_{i=1}^{n} W_i = 1$$

Définissons la médiane des e_i comme étant:

$$e^m = (e_m + e_{m'}) / 2$$

où $\sum_{K=1}^{m-1} W_K < 0.5 \leq \sum_{K=1}^{m} W_K$

et $\sum_{K=1}^{m'-1} W_K \leq 0.5 < \sum_{K=1}^{m'} W_K$

De façon similaire, le premier quartile e^{Q1} et le troisième quartile e^{Q3} sont définis de la façon suivante:

$$e^{Q_1} = (e_{Q_1} + e'_{Q_1}) / 2$$

où $\sum_{K=1}^{Q_1-1} W_K < 0.25 \leq \sum_{K=1}^{Q_1} W_K$

et $\sum_{K=1}^{Q'_1-1} W_K \leq 0.25 < \sum_{K=1}^{Q'_1} W_K$

et $e^{Q_3} = (e_{Q_3} + e_{Q'_3}) / 2$

où $\sum_{K=1}^{Q_3-1} W_K < 0.75 \leq \sum_{K=1}^{Q_3} W_K$

et $\sum_{K=1}^{Q'_3-1} W_K \leq 0.75 < \sum_{K=1}^{Q'_3} W_K$

Les valeurs de la médiane et des quartiles sont utilisées pour identifier les observations aberrantes. Pour ce faire, définissons les 2 distances suivantes:

$$d_{Q1} = \text{Max}\left[e^m - e^{Q1}, A * \text{abs}(e^m)\right]$$

et

$$d_{Q3} = \text{Max}\left[e^{Q3} - e^m, A * \text{abs}(e^m)\right]$$

où, A est une valeur arbitraire et abs (e^m) est la valeur absolue de e^m.

Le terme A*abs (e^m) est utilisé pour éliminer les problèmes qui surviennent lorsque $(e^m - e^{Q1})$ et/ou $(e^{Q3} - e^m)$ sont très petits.

Les unités pour lesquelles leur valeur de e est à l'extérieur de l'intervalle

$$(e^m - c\, d_{Q1},\ e^m + c\, d_{Q3})$$

sont considérées comme des observations aberrantes où c, est le nombre de "distances" inter-quartile (i.e. d_{Q1} ou d_{Q3}) à partir de la médiane.

Exemple de l'utilisation du terme A*abs(e^m)

Supposons qu'il y a 15 unités dans une cellule et que leur valeur de e sont les suivantes.

$e_1 = 78,\ e_2 = 79,\ e_3 = 80,\ e_4 = 98,\ e_5 = 98.5,$

$e_6 = 99,\ e_7 = 99.5,\ e_8 = 100,\ e_9 = 100.5,\ e_{10} = 101,$

$e_{11} = 101.5,\ e_{12} = 103,\ e_{13} = 125,\ e_{14} = 130$ et $e_{15} = 200.$

Le premier quartile est $e_4 = 98$, la médiane est $e_8 = 100$ et le troisième quartile est $e_{12} = 103$. Si le terme A*abs (e^m) n'est pas utilisé, les 2 distances sont:

$$d^{Q1} = e^m - e^{Q1} = 100 - 98 = 2$$

et $$d^{Q3} = e^{Q3} - e^m = 103 - 100 = 3$$

En utilisant une valeur de c = 8, on obtient les bornes suivantes:

1) Borne inférieure = $e^m - c\, d^{Q1} = 100 - 8*2 = 84$
2) Borne supérieure = $e^m + c\, d^{Q3} = 100 + 8*3 = 124$

En utilisant ces bornes, 6 unités sont identifiées comme observations aberrantes; ces unités sont:

$e_1 = 78$, $e_2 = 79$, $e_3 = 80$, $e_{13} = 125$, $e_{14} = 130$ et $e_{15} = 200$

Si une valeur de A = 0.0.5 est utilisé et que le terme A*abs (e^m) est employé, les 2 distances sont maintenant:

$$d^{Q1} = \text{Max}\left[e^m - e^{Q1},\ A * \text{abs}(e^m)\right] = \text{Max}(2, 5) = 5$$

$$d^{Q3} = \text{Max}\left[e^{Q3} - e^m,\ A * \text{abs}(e^m)\right] = \text{Max}(3, 5) = 5$$

En utilisant une valeur de c = 8, on obtient les bornes suivantes:

1) Borne inférieure = $e^m - c\, d^{Q1} = 100 - 8 * 5 = 60$

2) Borne supérieure = $e^m + c\, d^{Q3} = 100 + 8 * 5 = 140$

En utilisant ces bornes, uniquement 1 unité est identifiée comme observation aberrante. L'unité $e_{15} = 200$, qui est, en réalité, la seule observation aberrante.

Cet exemple montre bien que l'utilisation du terme A*abs (e^m) peut dans certaines circonstances, améliorer l'efficacité du système de vérification statistique.

3. Utilisation des paramètres A, c et u

Le paramètre "A" est utilisé pour éliminer les problèmes potentiels qui peuvent survenir lorsque la plupart des e_i sont regroupés autour de la médiane. L'utilisation de ce paramètre permet d'augmenter l'efficacité de la méthode de vérification statistique, en diminuant le nombre potentiel d'unités non-aberrantes identifiées par la procédure de vérification statistique.

Le paramètre "c" est utilisé pour contrôler l'étendue de l'intervalle d'acceptation. Si la valeur de "c" est augmentée, l'intervalle d'acceptation est plus grande et le nombre d'unités identifiées par la procédure de vérification statistique est diminué. Ce paramètre est très important et doit être utilisé pour contrôler le nombre d'unités identifiées comme observations aberrantes par le système de vérification statistique.

Le paramètre "u" est utilisé pour contrôler la forme de la courbe définissant les bornes inférieures et supérieures. Si la valeur de "u" est augmentée, moins de petites unités sont identifiées comme données aberrantes et plus de grosses unités sont identifiées comme données aberrantes. Ce paramètre est utilisé

pour modifier la courbe définissant les bornes de façon à ce qu'elles s'ajustent à la distribution des tendances.

4. Représentation graphique

L'intervalle d'acceptation généré par cette méthode de vérification statistique ainsi que l'effet des paramètres "c" et "u" sur cet intervalle peuvent être visualisés par l'intermédiaire des deux graphiques de la page suivante. Ces graphiques ont été produits de la façon suivante:

i) Pour chaque valeur de "c" et de "u" donnée, une borne inférieure (Min e) et une borne supérieure (Max e) pour les e_i ont été déterminées.

ii) Pour chaque observation i, les valeurs Min S_i et Max S_i ont été obtenues en appliquant la transformation inverse de E sur Min e et Max e respectivement.

iii) Pour chaque observation i, les valeurs Min R_i et Max R_i ont été obtenues en appliquant la transformation inverse de S sur Min S_i et Max S_i respectivement.

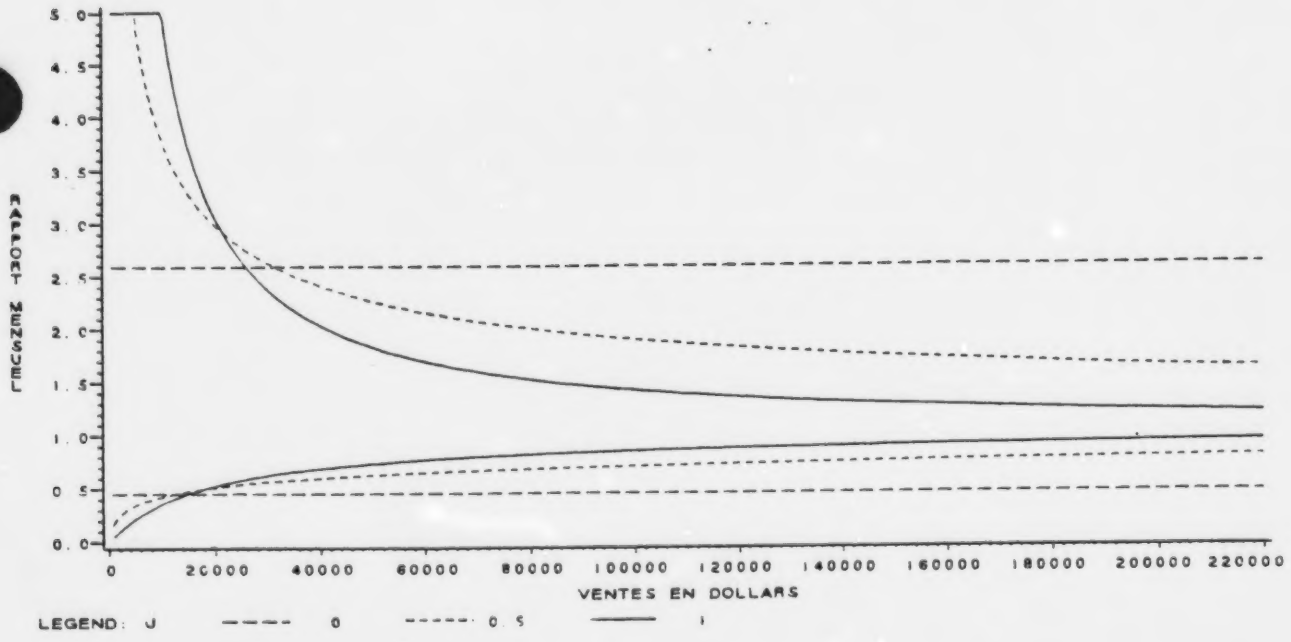
iv) Les courbes sur le graphique ont été produites en utilisant une interpollation entre chacun de Min R_i et Max R_i respectivement, afin de produire une courbe pour la borne inférieure et une courbe pour la borne supérieure.

Le premier graphique nous permet de voir l'effet du paramètre "u" sur la forme des courbes définissant la zone d'acceptation. Si la valeur de "u" est augmentée, la zone d'acceptation devient plus étroite pour les grosses unités et plus grande pour les petites unités.

Le deuxième graphique nous permet de voir l'effet du paramètre "c" sur les courbes définissant la zone d'acceptation. Si la valeur de "c" est augmentée, la zone d'acceptation devient plus grande, indépendamment de la grosseur des unités.

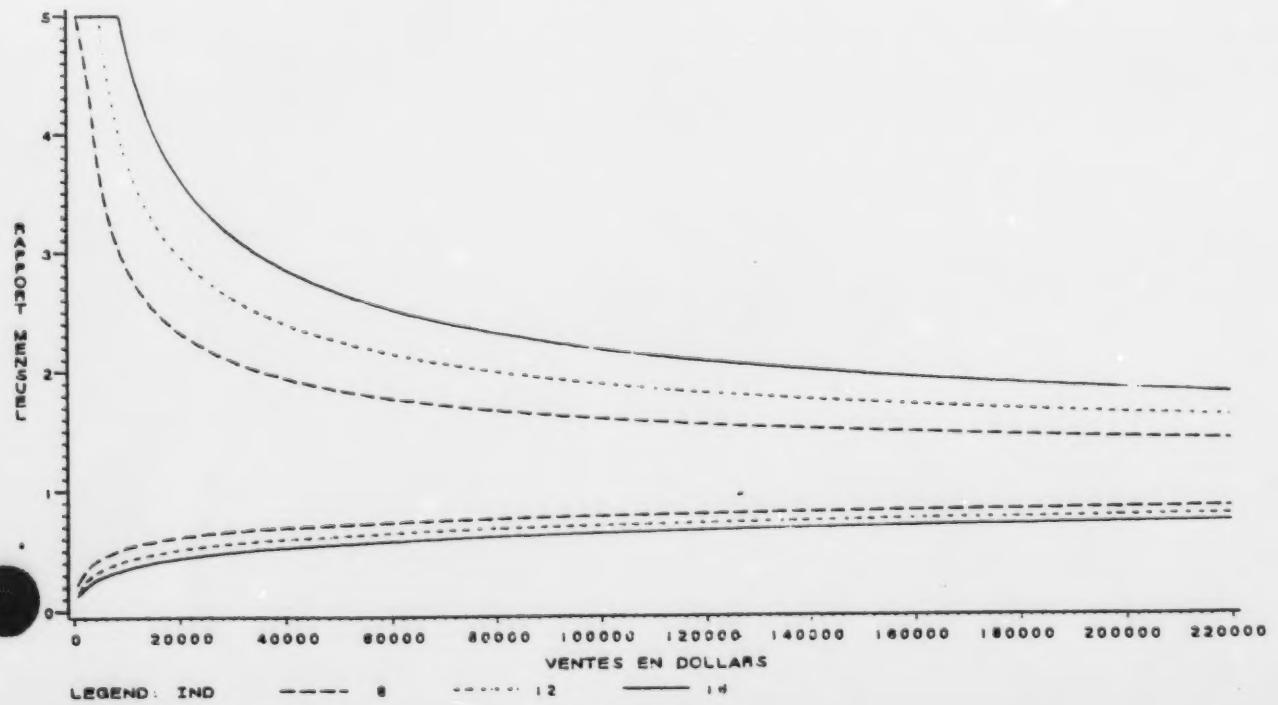
EDIT STATISTIQUE

C = 12 U = 0.0 , 0.5 , 27 1.0
SIC=647 TYPE=1



EDIT STATISTIQUE

U = 0.5 , C = 0 , 12 , 27 16
SIC=647 TYPE=1



5. Propriétés de la méthode proposée pour la vérification statistique

La méthode de vérification statistique proposé pour l'ECGCD a les propriétés suivantes:

i) Les bornes définissant les zones d'acceptation et de rejet des observations sont déterminées en utilisant l'information statistique des données de l'enquête, via l'utilisation d'une méthode de détection par les quartiles appliquée sur les "effets".

ii) Les bornes définissant les zones d'acceptation et de rejet varient en fonction de la taille des données financières et permettent une plus grande variation pour les unités de petite taille.

iii) La configuration des bornes peut être contrôlée par 2 paramètres. Le paramètre "u", qui contrôle la forme de la courbe définissant les bornes, et le paramètre "c", qui contrôle l'étendue de la zone d'acceptation.

6. Bibliographie

1. Berthelot, J.-M., Mars 1983, Spécifications du système de détection des données aberrantes pour l'Enquête sur le commerce de détail.

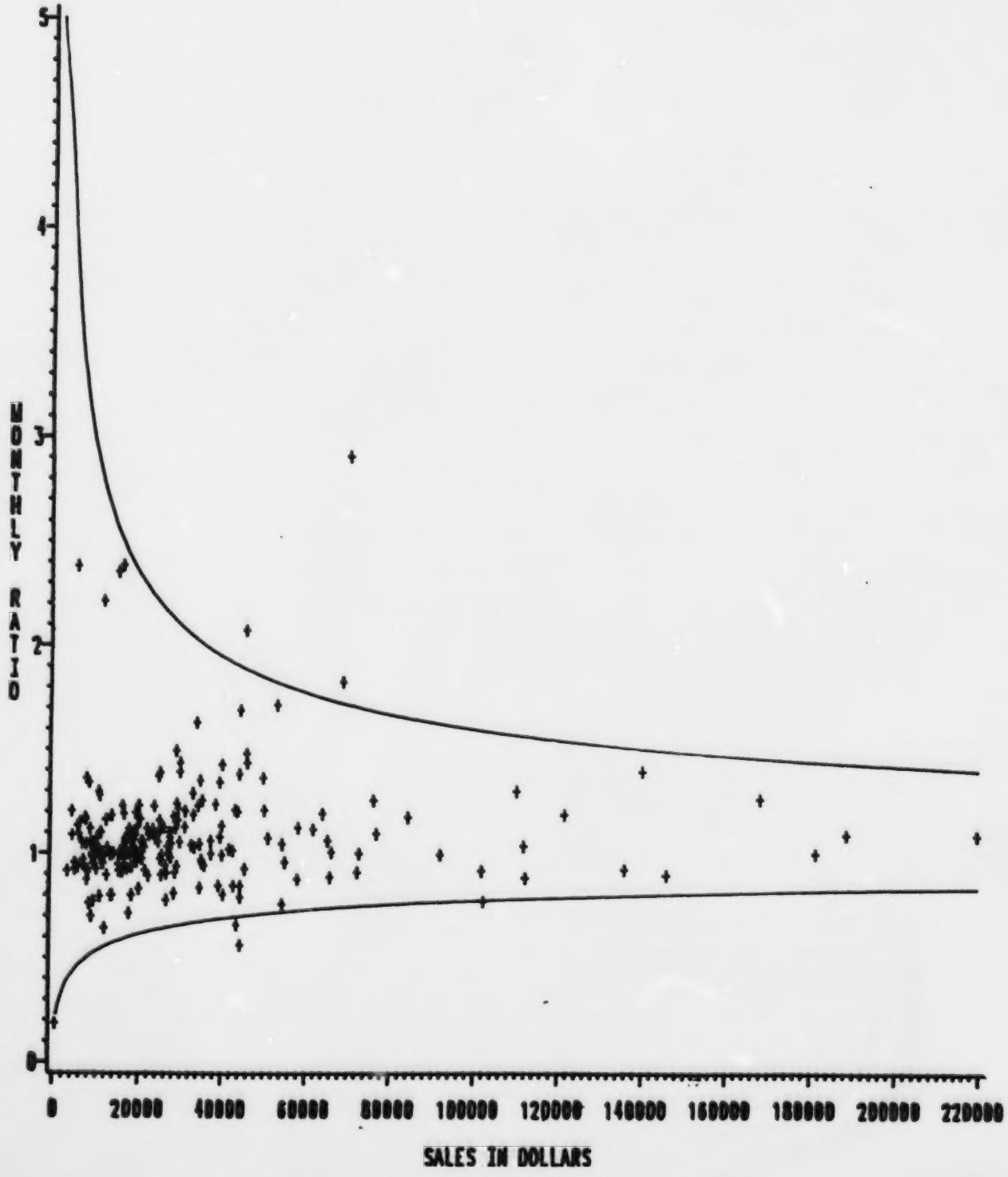
2. Hidiroglou, M.A., Avril 1983, Détection et traitement des observations aberrantes dans les sondages. Présentation faite à l'Université Laval.

3. Hidiroglou, M.A., Berthelot, J.-M., June 1982, Detection of outliers for Retail Trade.

4. Sande, I.G., September 1981, Estimation in the revised ISPI.

5. Sugavanam, R., December 1982, A statistical edit for change.

DIAGNOSTIC GRAPH SAMPLE
C = 8 , U = 0.5



Appendice I : Généralisation de la méthode de vérification statistique

La méthode de vérification statistique décrite dans les pages précédentes est restreinte, puisqu'elle a été développée dans le contexte de l'ECGD. Les restrictions suivantes ont été utilisées pour l'ECGD:

1) L'hypothèse que la médiane des tendances est près de 1 a été utilisée

2) La valeur de "U" a été restreinte sur l'intervalle [0, 1]

On peut généraliser cette méthode de vérification statistique en la définissant de la façon suivante:

Soient x et y, deux variables corrélées positivement

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, n paires d'observations pour les variables x et y.

r_i, le rapport entre y_i et x_i

et r_m, la médiane des r_i

Définissons la transformation "S", de la façon suivante:

$$S_i = \begin{cases} 1 - (r_m / r_i) & \text{si } r_i < r_m \\ (r_i / r_m) - 1 & \text{si } r_i \geq r_m \end{cases}$$

La transformation "E" est définie de la façon suivante :

$$e_i = S_c * [Max(x_i, y_i)]^u$$

où $-\infty \leq u \leq \infty$

La vérification statistique consiste à identifier les unités ayant une valeur de "e" aberrante, en utilisant la méthode de détection par les quantiles.

Note: Une valeur de "U" inférieure à zéro signifie que l'on permet moins de variation des x_i pour les unités ayant une petite valeur de x et y en comparaison avec les unités qui ont une grosse valeur de x et/ou y.

Ce système de vérification statistique généralisé a été testé en utilisant les données de l'Enquête Mensuelle sur le Commerce de Gros. Les trois graphiques des pages suivantes montent les résultats obtenus :

Graphique 1 : x = Ventes mois courant, y = Inventaires mois courant — E.C.C.

Graphique 2 : x = Ventes mois précédent, y = Ventes mois courant

Graphique 3 : x = Ventes annuelles, y = Ventes mois courant